IntoWeb : une plate forme hypertexte d'extraction de connaissances et de recherche d'information

Emmanuel Nauer

IntoWeb : une plate forme hypertexte d'extraction de connaissances et de recherche d'information

Emmanuel NAUER
Emmanuel.Nauer@loria.fr

LORIA - UMR 7503, Bâtiment B, Campus scientifique, B.P. 239, F-54506 Vandœuvre-lès-Nancy CEDEX, France.
Université Paul Verlaine – Metz, Île du Saulcy, B.P. 80794, 57012 Metz CEDEX, France.

Mots clefs :
Extraction de connaissances à partir de données, fouille de données, gestion de données, recherche d'information intelligente, web, système hypertexte
Keywords:
Knowledge extraction, data-mining, data management, intelligent information retrieval, web, hypertextual system
Palabras clave:
Extracción de conocimiento, minería de datos, gestión de datos, búsqueda de información inteligente, web, sistema hypertexto.

Résumé

Dans cet article, nous présentons un système hypertexte, nommé IntoWeb, qui fournit aux chercheurs ou spécialistes de l'information scientifique les moyens d'exploiter les données structurées sur leur domaine et des données – textuelles – du web pour des besoins de recherche d'information, d'analyse de leur domaine ou de veille. IntoWeb est un système générique d'exploitation de données qui implémente un processus complet et itératif d'extraction de connaissances à partir de données. Le système permet de manipuler différents types d'objets (documents structurés, documents textuels, vecteurs, classifications, etc.). Des opérateurs (génération d'un vecteur à partir d'un document textuel, classification de documents structurés, etc.) permettent d'exploiter chacun des différents types d'objets à des fins d'analyses ou de recherche d'information. L'application d'un opérateur sur un ensemble d'objets produit de nouveaux objets, à leur tour exploitable dans le système. La résolution complète d'un problème d'extraction de connaissances ou de recherche d'information prend la forme d'une succession d'opérations appliquées à des objets. Le choix des objets à exploiter et des opérations à appliquer à ces objets est à la charge de l'utilisateur et dépend du problème à résoudre ; l'enchaînement des opérations est grandement facilité par IntoWeb grâce à la mise en place d'une interface web simple à utiliser.

1 Introduction

La maîtrise de l'accès à l'information est primordiale dans de nombreux domaines, comme celui de la recherche ou celui de la veille scientifique et technique. Les données relatives à un domaine sont de plus en plus facilement accessibles, sur le web en particulier ; toutefois cette quantité croissante de données disponibles nécessite de mettre en œuvre des moyens particuliers pour les exploiter. Le but de nos travaux est de développer un environnement dans lequel chercheurs ou spécialistes de l'information scientifique peuvent exploiter les données structurées sur leur domaine et des données – textuelles – du web, pour des besoins de recherche d'information (RI), d'analyse de leur domaine ou de veille.

Nous présentons dans ce papier un système nommé IntoWeb, développé à cet escient. IntoWeb a été modélisé comme un système générique d'exploitation de données. Différents types d'objets y sont manipulables (documents structurés, documents textuels, vecteurs, classifications, etc.). Des opérateurs (génération d'un vecteur à partir d'un document textuel, classification de documents structurés, etc.) permettent d'exploiter chacun des différents types d'objets à des fins d'analyses ou de RI. L'application d'un opérateur sur un ensemble d'objets produit de nouveaux objets, à leur tour exploitable dans le système.

La section 2 pose le contexte général du travail et détaille les vocations majeures du système, la section 3 présente le système dans ses grandes lignes, la section 4 illustre des cas concrets d'utilisation ; nous concluons par les perspectives de notre travail.

2 Objectifs

La mise en œuvre du système IntoWeb a été motivée par les besoins de certains utilisateurs particuliers (chercheurs, experts, veilleurs) en terme de RI et d'analyse de domaines (analyses bibliométriques, en particulier) ; ces deux aspects étant très complémentaires [18]. La conception d'IntoWeb est une généralisation des principes majoritairement mis en œuvre dans les systèmes de RI ou d'analyses de corpus. Nous décrivons à présent ces principes.

2.1 Accès intelligent à l'information

Une approche générale qui couple l'exploitation de connaissances (extraites par des techniques de fouille de données) à un système de recherche d'information a été proposée dans [16] ; cette approche a permis de mettre en place un système opérationnel pour fouiller des données structurées (références bibliographiques) et accéder au web. Ce système, dénommé IntoBib, repose sur l'utilisation de l'hypertexte pour accéder de façon exploratoire aux données, dans le but d'identifier celles qui répondent à un certain besoin. Ces données peuvent alors être exploitées par des fonctionnalités de fouille (dénombrements, classifications, extractions de règles, etc.), déclenchées à la demande dans le but d'extraire de nouvelles connaissances capables de guider la recherche d'information. L'idée est que la fouille de données et la recherche d'information sont deux approches extrêmement complémentaires pour appréhender des données : la fouille de données permet de guider la recherche d'information à partir des connaissances extraites des données, et inversement, la recherche d'information permet de guider la fouille de données par l'exploitation des connaissances issues de la fouille de données [18]. Un enjeu du système IntoBib est également la production de connaissances qui donnent une idée synthétique du contenu de données structurées. Concrètement, cette production de connaissances peut se voir comme un processus d'extraction de connaissances à partir de bases de données (ECBD) qui agit sur des informations scientifiques et techniques (IST). Elle peut consister à chercher à dégager les principaux thèmes de recherche sous-jacents à un corpus de références bibliographiques, ou encore les collaborations entre auteurs, l'émergence d'une technique bien particulière, etc. Nous touchons en cela au domaine de la bibliométrie qui fixe les bases d'exploitation de l'IST. Les connaissances du domaine (réseaux d'auteurs, vocabulaire employé par un auteur, etc) peuvent alors être exploitées pour la RI sur le web. L'accès aux données du web est réalisé via un moteur de recherche classique (Google en l'occurrence) qui est utilisé comme un outil distant du

système. IntoBib fournit par conséquent un cadre général pour guider l'accès aux données du web, ceci en combinant l'accès par navigation hypertextuelle (par exemple dans des thèmes de plus en plus spécialisés) à l'interrogation d'un moteur de recherche, utilisé comme passerelles entre IntoBib et le web. L'utilisation de techniques d'ECBD permet dans ce contexte de déterminer automatiquement des requêtes ou encore d'assister l'utilisateur dans l'expression de son besoin ; cette aide à la formulation, par l'apport de connaissances extraites de données, permet d'améliorer considérablement l'efficacité de la recherche d'information sur le web [15, 2, 16][1]. D'une façon générale, l'exploitation de connaissances pour la recherche intelligente d'information est une approche qui a fait ses preuves [5, 11, 15, 12].

2.2 Extraction de connaissances à partir de données

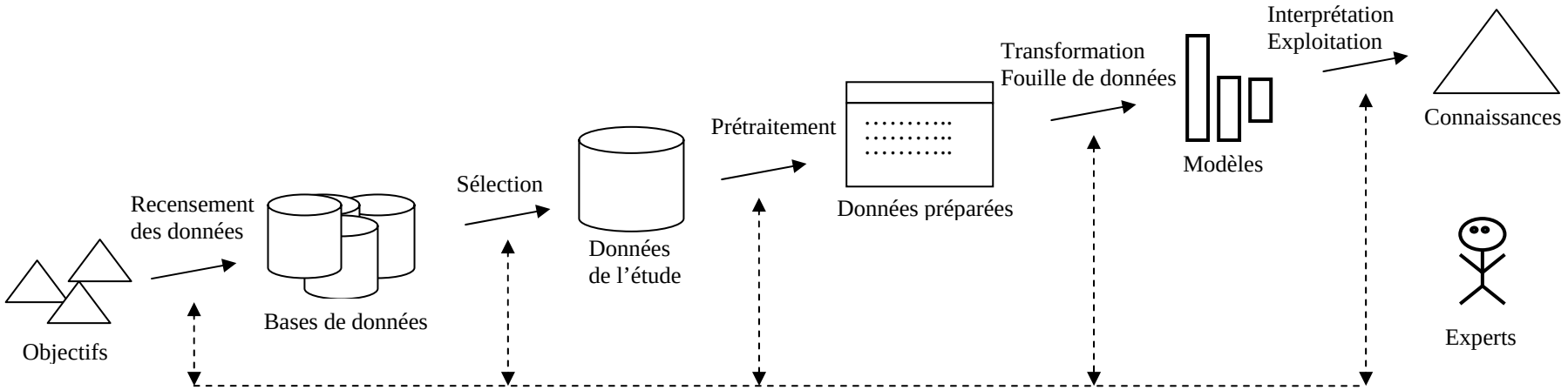

Figure 1. Les grandes étapes du processus d'ECBD

La conception d'IntoWeb repose sur une analyse des mécanismes couramment mis en œuvre dans le cadre d'un processus d'ECBD. La figure 1, adaptée de [8] synthétise les grandes étapes du processus d'ECBD. Après avoir défini les objectifs de l'étude (i.e facteurs à étudier), il s'agit de recenser les données disponibles ; les objectifs pourront être affinés, voire redéfinis en fonctions des propriétés des données recensées. Parmi les données disponibles, une sélection de données d'étude a lieu dans le but de répondre aux objectifs visés. Ces données sont ensuite prétraitées : les données bruitées (i.e. données erronées ou

[1] Une aide à la formulation à partir de connaissances extraites de données du domaine (pour interroger un moteur de recherche) s'avère insuffisante dans un processus complet – notamment itératif – de RI. L'utilisation d'un agent (ie. *crawler*) capable de parcourir le web de façon dirigée [4, 13] en suivant les liens hypertextes, tout en exploitant des connaissances (contexte, règles associatives, etc.) pour guider la RI et évaluer les documents rencontrés selon des critères propres est préconisé. Un tel outil est présenté dans [17].

incomplètes) doivent être supprimées ; une réduction des données en nombre (si leur nombre est important) et/ou en nombre de caractéristiques à conserver pour leur description peut également être réalisée. La méthode de fouille (algorithme et paramètres à utiliser) doit ensuite être sélectionnée pour être appliquée aux données afin de produire des modèles. L'interprétation des modèles par un spécialiste du domaine permet de juger si les éléments extraits par des méthodes de fouille sont pertinents, les érigeant ainsi au statut de connaissances. Le processus, idéalement chronologique, est en réalité itératif ; des retours en arrière sont souvent nécessaires pour ajuster les éléments de la chaîne complète. Le processus est également dirigé par des experts, expert(s) du domaine des données et expert(s) des méthodes.
Les processus d'analyse bibliométrique ou de veille sont des illustrations typiques de processus d'ECBD. Un parallèle peut également être fait avec le processus de recherche d'information dans lequel la prise en compte du résultat d'une recherche est souvent réalisée par un retour à l'étape de formulation du besoin, rendant ce processus également itératif. La similarité des processus mis en œuvre pour répondre à un besoin de recherche d'information, d'ECBD ou de veille, nous a conduit à une réflexion sur la modélisation d'un système générique d'exploitation de données. Ce système est une réification du processus présenté à la figure 1.

2.3 Modélisation générique d'un processus d'ECBD et de recherche d'information.

L'analyse des processus de recherche d'information, d'ECBD ou de veille a permis de dégager une approche générique de résolution de problème, qui consiste en une succession d'opérations élémentaires, appliquées à des objets de natures différentes. Une opération élémentaire consiste à transformer un ensemble d'objets en un autre ensemble d'objets, cet ensemble d'objets résultat pouvant à son tour être exploité par d'autres opérations. Par exemple, pour cartographier les activités d'un chercheur C, la première étape consiste à identifier les données relatives à C. Les **données brutes** sont hétérogènes et peuvent provenir de sources très diverses **: bases de données, documents textuels, web**, etc., ce qui nécessite de savoir intégrer, traiter et gérer ces différents types de données. Pour notre problème de cartographie, les données bibliographiques et/ou du web sont des données de départ classiquement exploitées. Une fois les **données intégrées** et traduites dans un format adéquat, les **opérations de fouille de données** peuvent être appliquées pour faire émerger des **éléments de connaissances** potentiellement exploitables dans un système intelligent. L'application d'opérateurs de dénombrement permet d'extraire les éléments dominants, la recherche de motifs fréquents – groupes de propriétés apparaissant dans les données avec une fréquence supérieure à un seuil donné – ou de règles permet d'extraire les régularités dans un ensemble de données, les méthodes de classification permettent de structurer un ensemble de données à travers une représentation condensée. Pour le problème de cartographie, une analyse de répartition en fréquence selon les années permet par exemple de juger de la productivité de C, une classification sur les mots-clés décrivant les travaux de C permet de dégager ses thématiques, l'extraction de motifs fréquents ou de règles peut par exemple permettre d'extraire, que dans un domaine particulier, C travaille toujours de façon conjointe avec les mêmes personnes. Les éléments de connaissances nouveaux produisent de **nouvelles données**, qui peuvent à leur tour être fouillées.
La figure 2, adaptée de [20], résume les interactions entre les différents types d'objets et d'opérations d'un système générique d'exploitation de données. Des objets sont sélectionnés en vue de l'application d'une certaine opération ; l'application de l'opération produit un résultat sous la forme d'un type d'objet également manipulable par le système. La résolution complète d'un problème se fait par un enchaînement d'opérations choisies par l'utilisateur sur des ensembles de données également choisis par l'utilisateur (le choix et la construction des ensembles de données résultent d'un processus de recherche d'information).

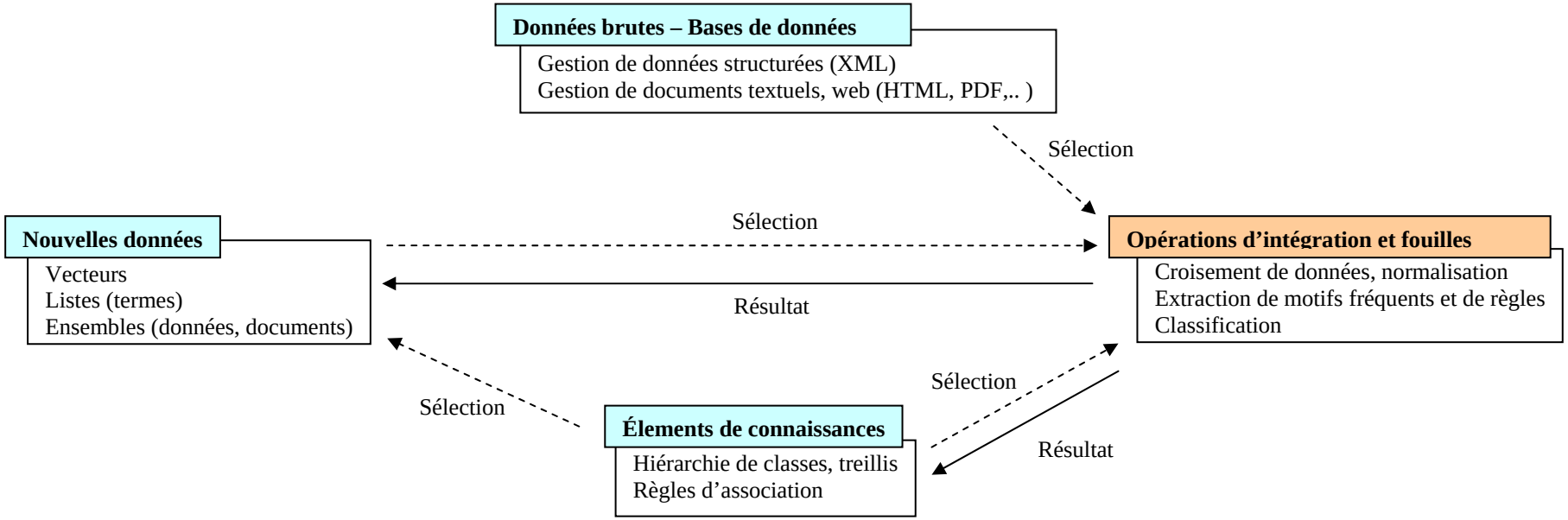

Figure 2. Exploitation de données dans le cadre d'un processus d'ECBD et de recherche d'information

3 IntoWeb

La construction d'un système générique d'extraction de connaissances nécessite tout d'abord de définir les différents types d'analyses souhaitées (en lien avec les objectifs des analyses statistiques, de fouilles de données ou de recherche d'information). Il s'agit ensuite de pouvoir opérationnaliser les processus mis en œuvre pour mener à bien ces analyses. Pour ce faire, il est nécessaire d'identifier précisément les types d'objets qui vont être manipulés ainsi que les opérateurs qui pourront exploiter ces différents types d'objets. Dans cette section, nous détaillons les différents types d'objets exploitables et les façons de les exploiter dans notre système IntoWeb. Nous exposons ensuite notre choix concernant l'utilisation d'un système hypertexte pour la mise en œuvre d'un système d'ECBD et de recherche d'information.

3.1 Types d'objets manipulés

Le système manipule 7 types d'objets. Parmi ces types, on peut distinguer ceux qui, initialement, alimentent le processus (objets en entrée) et les objets intermédiaires, produits en différentes étapes de l'expertise.

3.1.1 Types d'objets en entrée

Ces objets sont les données brutes sur lesquelles le processus d'exploitation va démarrer. On retrouve ici :

- des **données structurées** qui décrivent différents types d'information (références bibliographiques, brevets, etc.). Ces données doivent être représentées en XML [27], les informations considérées sont contenues dans des éléments XML de la forme `<propriété>valeur</propriété>`. Par exemple, le document XML présenté à la figure 3 introduit une référence bibliographique ; des chemins XPath [28] permettent de désigner de façon générique les propriétés à manipuler (par exemple : `doc/descripteur` désigne les mots-clés). Ces données proviennent en général de l'interrogation de base de données ; elles représentent le domaine à expertiser.

```
<doc>
   <auteur>E. Nauer</auteur>
   <titre>IntoWeb ...</titre>
   <descripteur>extraction de connaissances</descripteur>
   <descripteur>web</descripteur>
   ...
   <resume>Le besoin ...</resume>
</doc>
```

Figure 3. Exemple de données structurées au format XML

- des **documents textuels**, désignés par un URI, qui peuvent provenir du web ou d'une machine locale. Ces documents peuvent se présenter sous différents formats (.html, .doc, .pdf, etc.) mais seront toujours considérés comme du texte, à savoir une suite ou un ensemble de termes, selon l'exploitation qui en sera faite. L'objectif, dans la prise en compte du web est de pouvoir porter l'expertise à des données de nature, de fraîcheur et d'exhaustivité plus large que celles présentes dans les bases de données bibliographiques ou de brevets, par exemple.
- des **connaissances du domaine** : il s'agit de pouvoir manipuler des hiérarchies, qui peuvent être relativement sommaire (hiérarchie de termes, telle que celle d'un thesaurus par exemple) ou plus complexe, dans le cas d'une ontologie de domaine (codée par exemple en OWL [21]). L'objectif est ici de tirer profit de connaissances dans certaines phases de l'expertise qui est en cours.

3.1.2 Objets numériques

Ces objets sont issus de l'application de certaines opérations aux données d'entrée ou à des données produites en résultat à une étape de l'expertise. On retrouve ici des objets tout à fait classiques des systèmes de recherche d'information ou d'analyse de corpus tels que :

- des **vecteurs** : ensemble de termes pondérés (1 terme = 1 dimension de l'espace de représentation) permettant de représenter de façon synthétique l'information contenu dans un ou plusieurs document(s) XML ou textuel(s). L'utilisation du modèle vectoriel [25] permet de mettre en œuvre des calculs de similarité entre documents, un classement de pertinence de documents, un regroupement de documents en classes, etc.
- des **clusters** : organisation structurée composée d'un ensemble de classes et de leurs relations, résultant d'une classification numérique. Les classes peuvent être considérées comme un ensemble de valeurs agglomérées en raison de leur proximité. La mise en œuvre opérationnelle est l'implémentation

de l'algorithme du simple lien à partir de la cooccurrence de valeurs [14], mise en œuvre dans de nombreux outils tels que SAMPLER, SDOC, etc. L'idée est ici d'obtenir une représentation synthétique d'un ensemble de documents.

3.1.3 Objets symboliques

Ces objets résultent de l'application de méthodes de fouille de données symboliques aux données d'entrée ; les objets symboliques qui ont été intégrés sont :

- les **règles d'associations** : une règle d'association, notée P1 → P2 est *valide* si et seulement si tous les documents qui possèdent les propriétés de P1 possèdent également les propriétés de P2. L'extraction de règles d'association informatives, à savoir lorsque P2 $\not\subset$ P1, a été particulièrement étudiée dans [10]. En analyse bibliométrique, on peut identifier des règles du type « *tels descripteurs implique tels auteurs* », « *tels descripteurs et tels auteurs implique telle année* », etc.
- les **treillis** : structure hiérarchique, construite à partir d'une table de données « documents × propriétés », en agrégeant en classes les propriétés co-occurrentes et en organisant les classes selon une relation de spécialisation [7]. Comme pour les clusters, l'objectif est de construire une structuration hiérarchique des données pour en faciliter l'appréhension ; ce type de structure peut également être employée pour améliorer la recherche d'information par extension de requêtes [2] ou en réorganisant les résultats d'une recherche d'information [3].

3.2 Opérateurs

Nous détaillons maintenant les opérateurs applicables aux différents types d'objets, en illustrant chacun de ces opérateurs par une finalité d'expertise.

3.2.1 Données structurées

Les opérateurs applicables aux données structurées sont relativement classiques. Ainsi, il est possible :

- d'effectuer des **calculs ensemblistes** (intersection, union, différence) sur plusieurs ensembles de données afin de croiser, fusionner ou supprimer des données. Cette opération a plutôt vocation de recherche d'information ou d'une sélection affinée de documents dans l'ensemble des documents initiaux. Par exemple, extraire les références écrites par tel auteur sur telle thématique, une thématique pouvant par exemple être l'ensemble de mots-clés présents dans une classe issue d'une *clusterisation.*
- de réaliser une **analyse de répartition de fréquence** sur un type de propriété pour déceler les éléments dominants : par exemple les auteurs les plus productifs, les mots-clés les plus employés, les années phares pour une thématique de recherche ou une technologie, etc.
- déclencher une ***clusterisation*** sur un type de propriété (forcément multivaluée en raison de l'analyse de cooccurrence des valeurs) pour obtenir une représentation plus synthétique des éléments contenus dans les données structurées : par exemple, collaboration d'auteurs, thématiques dominantes, etc.
- d'extraire des **règles d'associations** multi-champs pour détecter les implications entre éléments : « tels mots-clés impliquent tels auteurs » qui exprime que tous les documents qui traitent d'un sujet ont été écrits par tels auteurs, « tels mots-clés impliquent tels mots-clés » exprimant que tous les documents traitant de A et B traitent également de C.
- d'en construire une **représentation synthétique sous la forme d'un vecteur** par extraction des valeurs associées aux propriétés à considérer ; la prise en compte de champs textuels (cas du titre ou du résumé d'une référence bibliographique par exemple) est traitée classiquement : extraction des mots, élimination des mots-vides, pondération [22].

Une orientation moins classique concerne l'exploitation des données structurées pour la recherche de documents sur le web. Ainsi, à partir d'un ensemble de documents locaux représentant un intérêt identifié par l'utilisateur, il est possible d'obtenir une aide à la formulation ainsi qu'une **formulation automatique**

de requête pour interroger le moteur de recherche Google et récupérer un ensemble de documents textuels du web. Des expérimentations concernant la formulation automatique de requêtes dans le cadre de la recherche d'information scientifique dans un domaine sont décrites dans [16] ; les approches expérimentées concernent une combinaison de termes utilisés dans les titres et résumés des références bibliographiques. L'utilisation du crawler décrit dans [17] permet de multiplier les requêtes initiales, de valider et d'ordonner les documents récupérés sur le web. La validation d'un document est réalisée en vérifiant la présence d'un certain nombre de termes connexes à ceux de la requête. On peut ainsi vérifier pour une recherche sur auteur, qu'au moins un de ses co-auteurs est présent, que son affiliation est présente, qu'au moins n termes utilisés dans les titres de ses publications est présents ? etc. Pour ordonner les documents, l'utilisation d'un vecteur relatif à des données locales similaires (pour un auteur, l'ensemble de ses références bibliographiques par exemple) permet de classer les documents valides.

3.2.2 Documents textuels du web

De nombreux opérateurs sont également disponibles sur les documents textuels. Ainsi, il est possible :

- de construire un **vecteur à partir d'un ensemble de documents**, dans le but d'en construire une représentation synthétique.
- d'extraire les **termes voisins** d'un ensemble de termes pour identifier les contextes d'apparition de certains termes dans les textes. Pour la recherche d'information, la consultation des termes à proximité des termes de recherche permet d'identifier le contexte dans lequel les termes de la requête sont utilisés et permet souvent de juger de la pertinence du document). Les termes voisins sont les termes qui apparaissent dans une fenêtre de n termes avant et m termes après un terme, n et m pouvant être fixés par l'utilisateur ; une suppression des mots-vides est également réalisée.
- d'extraire des **règles d'associations ou** déclencher une **clusterisation sur les termes voisins** d'un ensemble de termes, dans le but d'appréhender à un plus haut niveau les contextes dans lesquels certains termes apparaissent.
- d'établir un **classement de pertinence d'un ensemble de documents textuels par rapport à un vecteur**, l'objectif étant que l'utilisateur produise un vecteur qui *synthétise* son besoin, afin de l'utiliser, dans un cadre de RI, pour le classement de documents. Dans ce sens, un exemple typique d'enchaînement pour la recherche de documents sur le web en étant guidé par des données locales serait de produire un vecteur V à partir de documents structurés locaux, d'interroger le web (automatiquement ou non), puis de (re-)classer les documents retournés en réponse par Google en fonction de leur similarité avec V.

3.2.3 Vecteurs

Quelques opérateurs vectoriels sont à disposition pour permettre, à partir d'un vecteur :

- d'**extraire un sous-espace vectoriel** : l'objectif est de valider manuellement les éléments constituant le vecteur, par exemple, l'ensemble de termes qui décrits une thématique.
- d'**opérer des calculs vectoriels** : multiplication d'un vecteur par un cartésien, normalisation d'un vecteur, addition/soustraction de vecteurs. L'objectif est, par exemple, de produire un vecteur représentant au mieux le besoin l'utilisateur, ceci, à partir de documents jugés pertinents et d'autres jugés non pertinents [23], en combinant les vecteurs qui représentent chacun de ces deux ensembles.

De plus, comme mentionné précédemment en 3.2.2., un vecteur peut être utilisé pour classer des documents textuels.

3.2.4 Clusters, règles d'association et treillis

Ces trois types d'objets sont des types *complexes*, permettant d'obtenir une représentation de plus haut niveau des objets exploités en entrée. Nous ne souhaitons pas, dans les contextes d'applications que nous nous sommes fixés, pouvoir manipuler ces types d'objets autrement qu'en consultation. Cependant, pour pouvoir exploiter les résultats obtenus à travers ces représentations, il est important de pouvoir rattacher ces résultats aux objets d'entrée. Ainsi, la seule opération disponible sur ces types d'objets concerne l'accès aux documents desquels les éléments constitutifs du résultat ont été produits, à savoir l'accès aux documents :

- qui sont rattachés à un *cluster*,
- qui font qu'une règle d'association a été extraite (i.e. documents pour lesquels la règle est valide), ou
- qui sont rattachés à une classe du treillis.

3.3 Architecture fonctionnelle

Un des objectifs initiaux dans la mise en œuvre d'IntoWeb était de fournir une interface facile à appréhender, notamment par des utilisateurs non informaticiens. Pour cela, nous avons orienté notre choix vers une interface web, dont l'utilisation est relativement courante, pour n'importe quel utilisateur d'ordinateur. Ce type d'interface, et l'approche de navigation hypertextuelle qu'elle propose de façon naturelle, est particulièrement adapté à l'exploration de données, en proposant des types d'accès prédéfinis pour naviguer dans un corpus documentaire [6], ou ceux proposé dans des résultats produits par des outils d'analyse de l'information [9, 24]. L'émergence de technologies web (typiquement PHP, ASP, MySQL, etc.) a facilité l'accès en ligne à de nombreuses bases de données, incluant des opérations de plus haut niveau pour exploiter les données, à la volée. L'utilisation de ces techniques dans la mise en œuvre d'IntoWeb a conduit à un système très facile à utiliser. La construction du système a nécessité de mettre en place les fonctionnalités associées aux différentes étapes de *sélection*, de *traitement*, et de *visualisation* des résultats ; mais également de faciliter l'enchaînement de ces étapes. La consultation de données ainsi que la sélection d'un ensemble de données a lieu à partir de liens hypertextes ; le déclenchement d'une opération de traitement se fait en sélectionnant les données à exploiter et l'opérateur à appliquer. Le résultat obtenu est lui aussi explorable par navigation, des liens hypertextes permettent à nouveau de sélectionner des ensembles de données en vue d'une exploitation ultérieure. Ainsi, l'utilisateur n'est jamais limité dans l'exploitation et l'analyse de données ; le système permet d'effectuer une fouille effective des données et non une *pseudo-fouille* de résultats produits par des outils d'analyse d'information.
L'interface opérationnelle est composée de 3 fenêtres distinctes :

- une fenêtre de navigation qui permet d'accéder aux données brutes et d'y sélectionner des sous-ensembles pertinents pour le problème à résoudre ;
- une fenêtre de pilotage qui permet de gérer les sous-ensembles sélectionnés, et de leur appliquer des opérateurs d'analyse et de fouille (cette fenêtre retrace également l'historique des actions effectuées afin de pouvoir revenir à n'importe quelle étape du processus de fouille ou de recherche d'information) ;
- une fenêtre de résultats, dans laquelle sont affichés les résultats de l'application d'un opérateur à un ensemble d'objets ; il est alors possible de naviguer dans ces nouvelles données et d'y sélectionner des sous-ensembles en vue d'une exploitation ultérieure.

La section suivante montre comment IntoWeb permet de résoudre certains types de problème ; des copies d'écran sont données pour illustrer concrètement le processus de résolution.

4 Exemples d'utilisation de IntoWeb

Dans le cadre de différentes applications, IntoWeb a permis de montrer l'intérêt d'utiliser et de croiser des données structurées pour des besoins d'ECBD et de recherche d'information, ou tout simplement de disposer d'un système générique d'exploitation de données. Concrètement, IntoWeb a permis :

- de structurer un domaine pour un accès hiérarchisé à l'information : des accès thématiques sont construits automatiquement et de plus en plus finement par des méthodes de classification (à partir de descripteurs de références bibliographiques par exemple).
- de générer un environnement d'investigation spécialisé sur le web permettant à l'utilisateur d'être assisté dans l'étape consistant à définir le vocabulaire de la requête à soumettre à un moteur de recherche (pour une recherche d'information sur le web) ;
- de filtrer les documents en provenance du web : à partir des critères sélectionnés par l'utilisateur, une requête est générée automatiquement et est soumise au moteur de recherche Google. Cette requête ajoute un contexte de recherche (vocabulaire proche) aux critères sélectionnés. L'utilisation d'une représentation vectorielle du besoin permet également de classer les documents retournés en réponse.

Nous illustrons maintenant un peu plus en détail (1) comment peut se dérouler une session d'analyse de corpus, (2) comment utiliser des résultats d'analyse dans la recherche de documents web, pour finir par (3) l'interface de pilotage du système. Les étapes numérotées dans les figures 4 et 5 correspondent aux objets qui ont été concrètement sélectionnés puis exploités (cf. figure 6)

4.1 Analyse de corpus bibliographiques

La figure 4 illustre une partie d'une analyse bibliométrique. En partant des documents de l'auteur *« Emmanuel Nauer »* (étape 1), une clusterisation est construite sur les descripteurs, produisant un ensemble de classes qui sont les thématiques de recherche de l'auteur. La sélection d'une classe particulière, ici la classe traitant de *« ontologies – semantic similarity »* produit l'affichage du détail de la classe (étape 2). Les documents appartenant à cette classe (i.e. documents dont au moins un descripteur fait partie des descripteurs définissant la classe) peuvent être exploitées par n'importe quel opérateur sur les documents ; dans cette illustration, une nouvelle clusterisation est alors construite pour obtenir des thématiques plus spécifiques .

4.2 Recherche d'information sur le web guidé par les données

La figure 5, étape 5, présente une interrogation de Google pour récupérer les documents répondant à une certaine requête, ici *« nauer »*. Dans cette illustration, l'étape 4 représente le vecteur synthétisant les documents concernant *« Emmanuel Nauer »* (étape 1) dans le thème *« ontologies – semantic similarity »* (étape 2). Ce vecteur permet de reclasser les 100 documents de Google en fonction de leur proximité avec les documents concernant *« Emmanuel Nauer »* dans la thématique *« ontologies – semantic similarity »*. Les 6 premiers résultats présentés dans l'illustration étaient respectivement classés initialement en 4ème, 16ème, 58ème, 17ème, 15ème et 26ème position lors de l'interrogation Google ; le document concernant *« Emmanuel Nauer »* le plus mal classé se trouvant en 76ème position. Après reclassement des documents, il s'avère que tous les documents concernant*« Emmanuel Nauer »* sont classés en tête, de la 1ère à la 11ème place ; au-delà de la 11ème place, aucun des documents ne concerne *« Emmanuel Nauer »*. Un tri net est donc effectué entre les documents concernant *« Emmanuel Nauer »* et ceux ne le concernant pas.

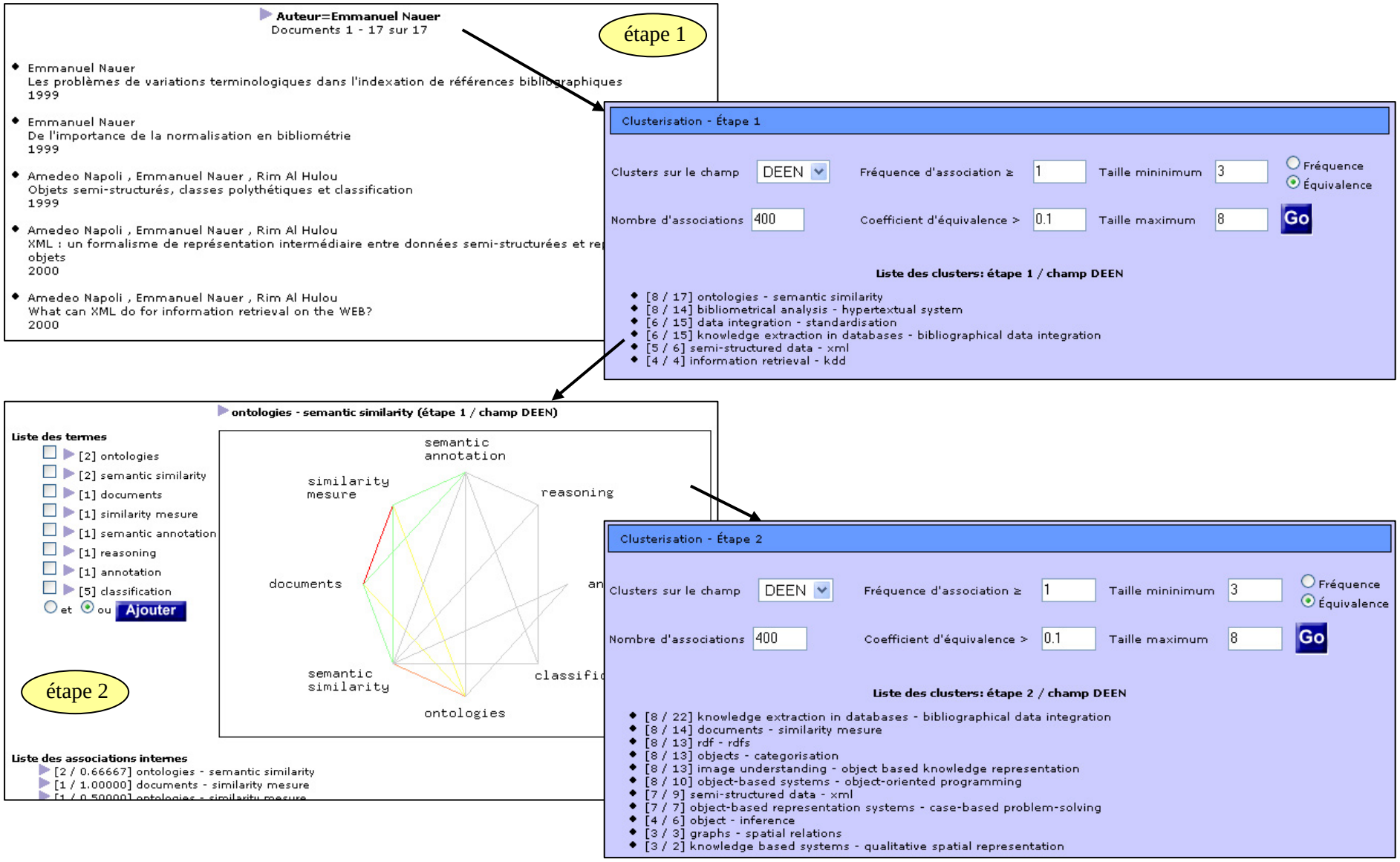

Figure 4. Illustrations de l'utilisation du système IntoWeb pour l'analyse de corpus bibliographiques

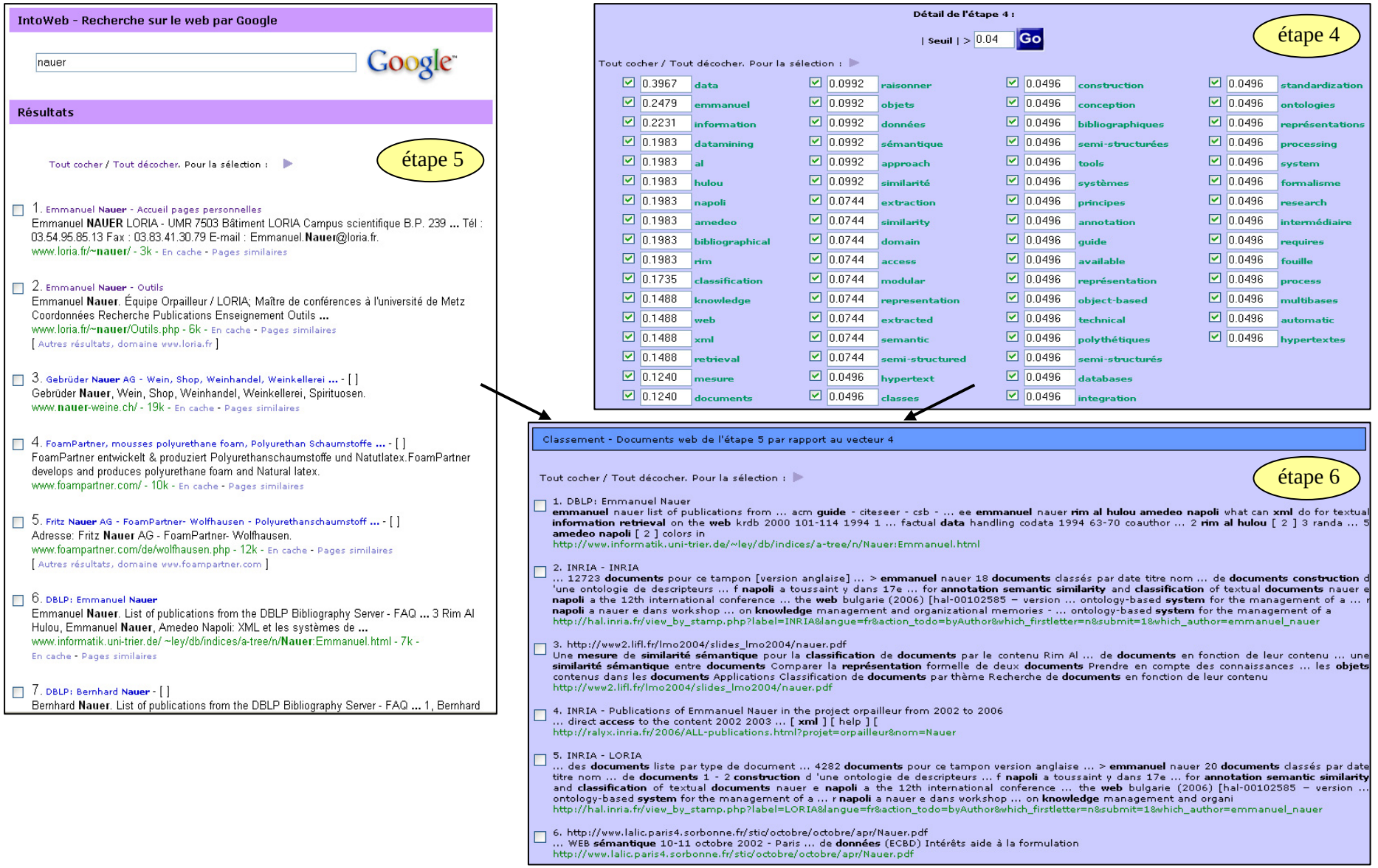

Figure 5. Illustrations de l'utilisation du système IntoWeb pour le classement de documents web

4.3 Interface de pilotage d'IntoWeb

La figure 6 présente l'interface de pilotage. La partie du haut liste les objets sélectionnés par l'utilisateur dans les différentes étapes d'exploitation des données, on y retrouve les objets utilisés dans les étapes données en illustrations dans les figure 4 et 5. La partie du bas liste les opérateurs disponibles, prêts à être déclenchés.

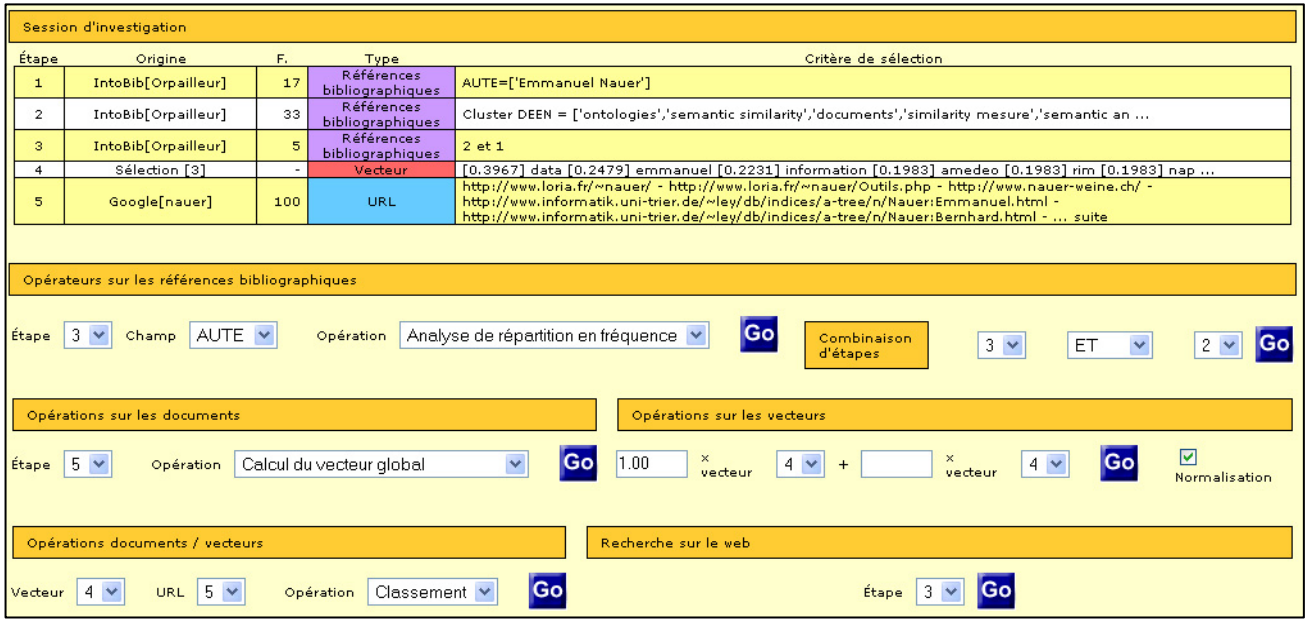

Figure 6. Interface de pilotage

5 Conclusion

Nous avons présenté un système hypertexte d'extraction de connaissances et de recherche d'information. Ce système permet de manipuler les objets les plus couramment utilisés dans les analyses de domaines et de recherche d'information. Les possibilités d'enchaîner les opérations élémentaires disponibles sont nombreuses et permettent de résoudre de nombreux types de problèmes de recherche d'information et d'ECBD. Les possibilités d'extension du système sont faciles à mettre en œuvre : de nouvelles opérations et/ou de nouveaux types d'objets peuvent être ajoutés ; l'appel à des outils externes permet d'étendre rapidement le système. Plusieurs directions d'extension sont envisagées ou déjà en cours d'intégration. Ces extensions concernent des perspectives de recherche liées au web sémantique [1], cadre dans lequel la mise en œuvre d'un système intelligent repose sur 3 points :

- les ontologies, qui de par la formalisation explicite des concepts d'un domaine, permettent une exploitation intelligente des données à travers des mécanismes d'inférence et de raisonnement ;
- les annotations, qui représentent le contenu des documents, construites par instanciation des concepts de l'ontologie ;
- les systèmes de traitement (raisonneur par exemple) capables d'exploiter les ontologies et les annotations pour résoudre des problèmes de façon intelligente (sur la base d'une sémantique formelle).

L'intérêt d'appliquer les idées du web sémantique à la recherche intelligente de documents bibliographiques est donné dans [26]. D'une façon plus générale, nous souhaitons intégrer la gestion d'ontologies (codée en OWL) et d'annotations RDF, types d'objets qui sont d'ors et déjà accessibles sur le web. Les ontologies permettent par exemple guider l'accès à l'information [12], l'exploitation conjointe d'annotations et d'une ontologie de domaine permet de manipuler des documents de façon intelligente par le contenu, comme nous l'avons déjà proposé dans [19].

Bibliographie

[1] BERNERS-LEE T., HENDLER J. and LASSILA 0. *The Semantic Web*, Scientific American, 2001.

[2] CARPINETO C., DE MORI R. and ROMANO G. Informative term selection for automatic query expansion. In *Proceedings of the Text REtrieval Conference (TREC-7)*, pages 363–370, 1998.

[3] CARPINETO C. and ROMANO G. Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO. In *Journal of Universal Computer Science*, 10(8), pages 985-1013, 2004.

[4] CHARKRABARTI S., van den BERG M. and DOM B. Focused crawling: A new approach to topic-specific Web resource discovery. In *8th World Wide Web Conference*, may 1999.

[5] CRAMPES M and RANWEZ S. Ontology-supported and ontology-driven conceptual navigation on the world wide web. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*, pages 191–199. ACM Press, 2000.

[6] DUCLOY J. DILIB, une plate-forme XML pour la génération de serveurs WWW et la veille scientifique et technique. Dans *MicroBulletin du CNRS*, 1999.

[7] DUQUENNE V. Latticial structures in data analysis. In *Theoretical Computer Science*, 217, pages 407–436, 1999.

[8] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P. and UTHURUSAMY R. *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.

[9] GRIVEL L. *L'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'information scientifique et technique*. Thèse en Sciences de l'Information et de la Communication, Université de droit et des sciences d'Aix-Marseille, 2000.

[10] GUIGUES J.L. Familles minimales d'implications informatives résultant d'un tableau de données binaires. In *Mathématiques, Informatique et Sciences Humaines*, pages 5–18, 1986.

[11] McGUINNESS D.L. Ontological Issues for Knowledge-Enhanced Search. In GUARINO N., editor, *Formal Ontology in Information Systems, Proceedings of the 1st International Conference (FOIS'98)*, pages 302–316, Trento, Italy, 1998. IOS Press.

[12] HERNANDEZ N. et MOTHE J. Ontologies pour l'aide à l'exploration d'une collection de documents. Dans *Veille Stratégique Scientifique et Technique – VSST'2004*,2, pages 405–416, 2004.

[13] MAEDCHE A., EHRIG M., HANDSCHUH S., STOJANOVIC L. and VOLZ R. Ontology-focused crawling of documents and metadata.In *Proceedings of the 11th International WWW Conference*, 2002.

[14] MICHELET B. *L'analyse des associations*. Thèse en informatique, Université Paris VII, 1988.

[15] NAUER E., DUCLOY J and LAMIREL J.-C. Using of multiple data source for information filtering : first approaches in the MedExplore project. In *5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, 1997.

[16] NAUER E. *Principes de conception de systèmes hypertextes pour la fouille de données bibliographiques multibases*. Thèse en informatique, Université Henri Poincaré Nancy 1, 2001.

[17] NAUER E. DefineCrawler : un crawler paramétrable pour la recherche d'information intelligente sur le Web. Dans *Journées scientifiques Web sémantique*, 2002.

[18] Nauer E. Complémentarité entre fouille de données et recherche d'information dans le cadre d'analyses bibliométriques. Dans *13ème Congrès francophone AFRIF-AFIA de Reconnaissances des Formes et d'intelligence Artificielle*, 3 (2002), pages 965–974, 2002.

[19] NAUER E. and NAPOLI A. A proposal for annotation, semantic similarity and classification of textual documents. In *12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2006). AI, people and the web*, Bulgaria, 2006.

[20] *Knowledge Discovery guided by Domain Knowledge*, Orpailleur, 2006 research project activity report. http://www.inria.fr/rapportsactivite/RA2006/orpailleur/orpailleur_tf.html

[21] OWL Web Ontology Language Reference, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/owl-ref.

[22] VAN RIJSBERGEN C. J. *Information Retrieval*, Butterworths, 1979.

[23] ROCCHIO Jr. J.J.Relevance feedback in information retrieval. In *The Smart System -- Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971.

[24] Cristelle ROUX C, SOSSON D et DOUSSET B. XPlor : un outil d'investigation en ligne sur des données relationnelles. Dans *Veille Stratégique Scientifique et Technique – VSST'2004*, 2, pages 295–306, 2004.

[25] SALTON G.and Mc GILL M. J. *Introduction to Modern Information Retrieval*, Mc Graw-Hill, 1983.

[26] AL-SUDANI S., ALHULOU R., NAPOLI A. and NAUER E. OntoBib: an Ontology-Based System for the Management of a bibliography. In *Workshop on Knowledge Management and Organizational Memories - ECAI'2006*, 2006.

[27] Extensible Markup Language (XML) 1.1 (Second Edition), W3C Recommendation, 16 August 2006, http://www.w3.org/TR/xml11.

[28] XML Path Language (XPath) 1.0, W3C Recommendation, 16 November 1999, http://www.w3.org/TR/xpath.